

CASE STUDY

How the World's Largest Winery Built an AI Stack That Ages Well with Shakudo

4x

faster engineering delivery after compressing sprint cycles to twice weekly

INSTANT

deployments, down from four weeks, with no infrastructure handoffs

70M+

cases shipped a year, on a supply chain now backed by an AI gateway

PROFILE

Founded in 1933 in California, GALLO is the world's largest family-owned winery and top global wine producer. Evolved into a total-beverage giant (wine, spirits, beer, RTDs), its portfolio includes Barefoot, High Noon, and Four Roses. GALLO ships 70+ million cases annually, running a highly sophisticated global supply chain.

WATCH THE VIDEO

GALLO is the largest family-owned winery in the world and the largest wine producer by volume globally. Founded in 1933 and headquartered in Modesto, California, GALLO has grown into a full total-beverage-alcohol company spanning wine, spirits, beer, and ready-to-drink categories. Its portfolio includes household names such as Barefoot, New Amsterdam Vodka, E&J Brandy, High Noon, and Montucky, alongside its recent \$750 million acquisition of Four Roses Bourbon. GALLO ships upwards of 70 million 9-liter cases of alcohol per year and moves roughly half a million cases per day across its U.S. logistics network alone, making it one of the most sophisticated manufacturing and supply chain operations in the consumer goods industry.

For most of the last 25 years, the bottleneck inside large enterprises was writing software. Provisioning hardware was slow, but engineering capacity was slower. The arrival of mature coding assistants in 2024 inverted that equation almost overnight.

THE PROBLEM: CODE SHIPS IN DAYS, INFRASTRUCTURE CATCHES UP IN WEEKS

At GALLO, the shift became concrete when IT leadership began vibe-coding internal applications over a weekend and producing working software in two or three days. The traditional deployment pipeline, designed for a winery rather than a hyperscale software company, then took roughly four weeks to get that same code live.

That gap is the modern enterprise problem in a single sentence. Generative AI has collapsed the cost of producing production-grade code, and the constraint has migrated back into the layer it occupied in the early 2000s: infrastructure, environments,

CI/CD discipline, identity, access, and the orchestration of an exploding ecosystem of tools.

"My ability to ship code is faster than the ability to deploy it. Infrastructure is now the bottleneck."

Robert Barrios

Chief Information Officer, GALLO

Layered on top of the deployment gap were three additional pressures that any large, regulated, manufacturing-led enterprise will recognize. First, the economic logic of the traditional SaaS stack has weakened, because every major vendor is optimizing to preserve or grow ARR while AI-native development makes in-house CRMs, ERPs, and operational tools genuinely viable for forward-thinking IT organizations. Second, the rise of agentic systems introduced a new operating cost called token economics, where a single inefficient context window or a heavyweight MCP server can multiply spend by hundreds of turns inside one agent run. Third, manufacturing and supply chain decisions, such as least-cost shipping algorithms and production scheduling, require deterministic, auditable outcomes that a raw LLM cannot provide on its own.

GALLO also operates inside a category defined by trade secrets, regulated distribution, and proprietary formulations. Any platform that touches its data has to keep that data inside GALLO's governance boundary, with platform-wide audit trails, lineage, and network policies, while still giving builders access to the best frontier and open source AI tools available on any given day.

WHY GALLO CHOSE SHAKUDO

GALLO evaluated Shakudo against the specific shape of the problem above rather than against a generic platform checklist. Three drivers stood out.

The first was deployment velocity for microservices. The CIO organization needed an environment where an application coded over a weekend could be promoted into a governed production runtime in the same week, with CI/CD, observability, and access control attached by default rather than reassembled by hand each time. Shakudo provides that runtime inside GALLO's own cloud footprint, which removes both the multi-week deployment tail and the data residency questions that come with sending workloads to external SaaS environments.

The second was unified access to data and models for a broader population of builders. GALLO's vision extends beyond its central engineering team to technically capable people sitting inside business functions who can build applications when given safe access to the semantic layer, the data warehouse, and reusable APIs. Shakudo's tool-

agnostic orchestration of long-term storage, real-time stores, identity, access control, and secret management gave GALLO a single substrate for that broader builder population without forcing a single-vendor lock-in across the AI and data stack.

The third was hardware and workplace economics. Because Shakudo runs the IDE, the agent harness, the model endpoints, and the compute in the cloud, GALLO can decouple local hardware limitations from engineering velocity, allowing developers to drive frontier coding agents at full power using standard, lightweight workstations.

"What drew me in is simple. When developers ship production-ready code this quickly, how can I have environments spun up fast enough? Shakudo is how we close that gap."

Robert Barrios

Chief Information Officer, GALLO

THE SOLUTION: AN OPERATING SYSTEM FOR AGENTIC DEVELOPMENT AT GALLO

With Shakudo in place, GALLO has begun rebuilding its internal application portfolio around an agentic operating model rather than around traditional SaaS seats. The pattern is consistent across projects.

A product owner, often technical, partners directly with the business to define what winning looks like for a given workflow. Coding agents, including Claude Code style harnesses, then produce candidate implementations at very high cadence. Shakudo handles the surrounding concerns that historically consumed weeks of effort: provisioning the runtime, wiring identity and secrets, exposing the right data sources through the semantic layer and APIs, attaching logging and monitoring, and routing the resulting microservice through CI/CD into a governed environment. The four-week deployment tail collapses into something measured in hours.

The same platform supports a deliberately heterogeneous, multi-model strategy. Inside a single agentic ecosystem, GALLO routes tasks dynamically based on complexity: frontier-class models handle high-reasoning product owner and QA roles where logic is decisive, mid-tier models handle architecture and core development, and highly optimized, lightweight models handle routing and orchestration.

Because Shakudo treats models as interchangeable building blocks rather than a single-vendor commitment, GALLO can seamlessly swap in self-hosted, open-weight models for routine workloads the moment they achieve parity with commercial frontier alternatives.

"I do not want to be in a position where I have to pay for a token for every single piece of work. Eventually I want to buy compute, scale it, and run our own LLMs next to our data."

Robert Barrios

Chief Information Officer, GALLO

This is precisely the architecture Shakudo is built to support. Sensitive data stays inside GALLO's governance boundary. Audit trails, lineage, and network policies provide the virtual air-gap posture that a regulated, trade-secret-heavy beverage manufacturer requires when placing advanced models next to proprietary supply chain, formulation, and logistics data. Tool choice remains open, so GALLO's team can adopt new agent frameworks, skill specifications, hook patterns, and MCP servers as the ecosystem evolves on a five to six month innovation cycle without re-engineering the underlying platform each time.

DETERMINISM, TOKEN ECONOMICS, AND THE NEXT OPERATING MODEL

Two themes increasingly define GALLO's engineering culture on top of Shakudo. The first is the pursuit of determinism inside fundamentally probabilistic systems. By combining well-scoped agents, the emerging Agent Skills specification, and hooks for hard guarantees, GALLO is targeting outcomes in the 90 to 95 percent probabilistic range for general agent behavior, with hooks pushing critical paths toward full determinism. That matters most in supply chain, manufacturing scheduling, and logistics optimization, where the algorithm behind a decision directly affects how product reaches consumers and where the business will continue to validate outputs side by side with the system.

The second theme is token economics as a first-class engineering concern. Heavyweight MCP servers can consume tens of thousands of tokens before any useful work begins, and large files pulled into context early in a 300-turn agent run are dragged across every subsequent turn. GALLO's approach, enabled by Shakudo's flexibility in choosing harnesses, models, and runtime patterns, is to favor on-demand skill discovery over heavy MCP loading, to route work to the cheapest sufficient model per agent role, and to plan for self-hosted inference as open source quality continues to climb.

*"Token economics is going to be a thing. Determinism is going to be a thing.
Whoever designs for both wins the next decade of enterprise AI."*

Robert Barrios

Chief Information Officer, GALLO

Underneath these themes is a broader bet on what the enterprise operating model becomes in the age of AI. Transactional data-entry UIs give way to conversational and report-driven interfaces. Departments organized around five things done well map naturally onto agent swarms organized around the same scopes. Player coaches and outcome owners replace pure management layers. None of this works without a platform that can absorb constant change in models, frameworks, and tooling while keeping data, identity, and governance stable.

LOOKING FORWARD

GALLO's roadmap on Shakudo extends naturally from where the company is today. More microservices will be built and deployed by a wider population of technical builders across the business. More workloads will shift onto self-hosted open-source models running on GALLO's own compute via the Shakudo Platform, with frontier closed models reserved for the highest-leverage reasoning tasks. More agent swarms will take on supply chain, logistics, and manufacturing scheduling workflows, with hooks and human review designed into the critical decision points from day one.

For other large enterprises in critical infrastructure industries, the pattern is repeatable. The bottleneck has moved from writing code to deploying it safely, governing it correctly, and operating it economically next to proprietary data. Shakudo is the operating system for data and AI that closes that gap inside your own cloud or on-prem environment, with deep controls for audit, lineage, identity, and network policy already in place. Kaji is the autonomous AI agent that runs on top of it, connected to your data, equipped with the best open and closed source tools, and governed end to end through the Shakudo AI gateway so that employee and agent activity sits under a single immutable audit trail.

If your organization is staring at the same gap GALLO identified, where code ships in days and infrastructure still takes weeks, the next step is straightforward. **Get a demo of Shakudo and Kaji today** and see what the next enterprise operating model looks like running inside your own four walls.

GET IN TOUCH WITH US

Ethan Brooke

Senior Account Executive

ethan@shakudo.io

(437) 344-4349